Customer Segmentation and Analysis Based on Gaussian Mixture Model Algorithm

Eka Angga Laksana[1] and Marchel Maulana Fahrezi [2]

[1,2]Widyatama University, Bandung, Indonesia
[1]Eka.angga@widyatama.ac.id, [2] marchel.fahrezi@widyatama.ac.id

Abstract. The digital age has changed the business paradigm with digital marketing becoming a key element in dealing with modern market dynamics. Changes in consumer behavior in online content consumption encourage companies to utilize digital technology to reach a wider audience and connect personally. A deep understanding of consumer buying behavior is essential, enabling companies to design responsive and relevant marketing strategies. This research also highlights the importance of segmenting customer buying behavior in the face of intense competition. Through clustering analysis using the Gaussian Mixture Model (GMM) algorithm, consumer spending data is reduced and grouped into clusters that allow companies to understand consumer preferences and tendencies. The experiment shows that there are 4 optimal cluster as basic information for further analysis. Each cluster leads to marketing strategies, such as emphasis on health and active lifestyles, increased sales of specific products, and education of low-spending clusters. This analysis also emphasizes the importance of data preprocessing and feature selection in ensuring the accuracy of clustering results.

Keywords: clustering analysis, Gaussian Mixture Model, marketing strategies

1 Introduction

The digital era continues to evolve, digital marketing is not only a business strategy, but has become an indispensable key element in dealing with modern market dynamics [1]. Changes in consumer behavior, especially in terms of online content consumption, have encouraged companies to leverage digital technology to reach a wider audience and connect more personally [2]. Understanding how consumers purchase behavior has become a crucial element in effectively running business operations in a rapidly changing market. The term “customer buying behavior” refers to the diverse activities, thoughts, and emotions that buyers experience when choosing a product [3]. This complex and dynamic process is influenced by a few factors, both internal and external, such as personal beliefs, social standards, marketing, and product characteristics [4]. The many factors that influence the buying process make a deep understanding of consumer behavior essential for the success of companies during intense competition.

V. Mardiansyah and B. A. Prasetyo (eds.), *Proceedings of the Widyatama International Conference on Engineering 2024 (WICOENG 2024)*, Advances in Engineering Research 252,
https://doi.org/10.2991/978-94-6463-618-5_8

In fact, a mature understanding of these dynamics allows companies to design marketing strategies that are not only responsive but also relevant to the rapidly changing needs of the market. The success of digital marketing lies not only in reaching a wider audience, but also in the company's ability to establish a more personalized connection with consumers [5]. This era demands more than just product promotion, companies need to create engaging and relevant experiences for their consumers. Therefore, holistically understanding the term “customer buying behavior” is key for companies that want to build solid and sustainable relationships with their customers [1]. Segmentation and strategy customization are becoming increasingly important to meet the increasingly diverse needs of various consumer segments. In this context, a marketing paradigm shift is becoming increasingly important. It is not only about how products are marketed, but also about how companies understand and respond to the needs and desires of customers in various segments. By knowing the segmentation of customer purchasing behavior, companies can make marketing approaches based on consumer preferences [6].

In the competitive business world in the research that has been done, the advantages that can be the key for a company to be able to survive in the competitive world of companies, namely being able to understand consumer tastes and needs and being able to provide more satisfaction than what is offered. Digital marketing approaches are more effective through segmentation analysis of purchasing behavior, companies can determine different preferences, needs, and motivations among certain consumer groups [7]. For example, a group of consumers who already have a family may prefer health-focused marketing, while a group of consumers who do not have a family may be more interested in things that are considered luxurious. By better understanding customer buying behavior, companies can design more targeted marketing campaigns.

One of the approaches that this research focuses on for segmenting consumer purchasing behavior is the utilization of clustering techniques. Clustering, as a modern data analysis method, plays a crucial role in breaking down datasets into homogeneous groups based on similar characteristics [8]. In the context of dynamic consumer purchasing behavior, clustering techniques not only provide an in-depth understanding of patterns that may be missed directly, but also provide a foundation for thoroughly exploring the variety and complexity of consumer behavior. Companies can apply clustering techniques to group consumers based on diverse criteria, ranging from purchase patterns, purchase frequency, to brand preferences [9]. The results of clustering analysis not only provide comprehensive insights into the needs and motivations of each consumer group, but also open the opportunities to customize marketing strategies with a higher degree of precision. The application of data artificial intelligence in the clustering process opens the door to deeper understanding. By collecting and analyzing data thoroughly, companies can identify complex patterns, uncover hidden variables, and respond to trends that may go undetected with traditional methods. This innovation can help companies to design marketing strategies that are more adaptive and responsive to the changing dynamics of consumer behavior [10].

The solution of using clustering in consumer buying behavior research is not only an innovative move, but also a long-term investment to maintain a competitive advantage. In the era of ever-evolving digital marketing, companies that can deeply understand and quickly respond to the complex dynamics of consumer behavior will have an advantage in creating more personalized, relevant, and satisfying experiences for each consumer segment [11]. Thus, the application of clustering technology not only provides a deeper understanding of consumer behavior, but also allows companies to proactively adjust their marketing approach. Through this approach, companies are expected to build stronger relationships with consumers, create loyalty, and hopefully win the competition in an ever-changing market.

2 Methodology

This research has several steps starting from the preliminary stage to the evaluation stage. In general, this research will be divided into 3 major stages, namely: (1) Data Collection Stage, (2) Data Preparation Stage, and (3) Clustering Stage. The research stages in are explained as follows: The data collection stage begins through literature studies on previous research relevant to this topic. The next stage is data preparation before the data will be clustering, and the last stage is clustering to draw conclusions on the clustering results. Customer segmentation helps in understanding the needs and preferences of different customer groups. By dividing the market into smaller segments, companies can tailor their marketing, sales, and customer service strategies more effectively. This enables better focus on the specific characteristics of each segment, allowing companies to provide better added value and build stronger relationships with customers. The second stage is understanding and selecting the dataset that will be the topic of research. This stage is necessary to understand the problem that needs to be solved and set the goals to be achieved. Third, preprocessing is an important stage in data processing to improve the quality and sustainability of the analysis. At this stage, the data that has been obtained will undergo a series of transformations. Some common steps in preprocessing involve data cleaning from noise, normalization to scale the data, and feature extraction to detail relevant information. Data cleaning involves identifying and handling invalid, missing, or ambiguous data. Normalization helps in ensuring that the data has a uniform scale, so that no attribute dominates the other. Fourth, data exploration stage, data analysis is presented in the form of visualizations to easily understand data, correlations, and others. In this stage, several features are added such as age which is obtained by the customer's date of birth column, the marketing acceptance feature which sums up all ‘AcceptedCmp’ columns and the ‘Response’ column as a representative of whether the customer accepts marketing offers or not at all.

a) Finding the Optimal Number of Clusters

Finding the optimal number of clusters using the Bayesian Information Criterion (BIC) algorithm, Bayesian Information Criterion (BIC) is a statistical model evaluation method used to select the best model from a few alternative models [12]. It combines

two main components: model fit and model complexity. BIC measures the extent to which the model can explain the observed data. The better the model explains the data, the higher the model fit value. In other words, BIC favors the model that provides the most accurate representation of the data. Model complexity is related to the number of parameters used in the model. BIC penalizes model complexity by decreasing the model score as the number of parameters increases. This helps prevent overfitting, where the model over-adapts to the training data and loses its general ability to generalize to new data.

BIC can be formulated as follows:

$$BIC = -2 * \log(L) + k * \log(n) \qquad (1)$$

Where:
- L is the maximum likelihood of the model against the data.
- k is the number of parameters in the model.
- n is the number of data samples.

The best model selection is done by choosing the model that produces the lowest BIC value. A lower BIC value indicates a good balance between model fit and model complexity, so the model is considered better in the context of statistical modeling. BIC is often used in selecting the optimal number of clusters in cluster analysis.

The clustering process is carried out using the Gaussian Mixture Model (GMM) method with the number K which has been obtained from the Bayesian Information Criterion (BIC) results. GMM is a clustering algorithm that can identify data groups with different Gaussian distributions [13]. The use of K as the number of clusters that have been optimized from BIC helps ensure that the results and the number of clusters to be created are the best results. After obtaining the clustering results, the next step is to explore the data from each cluster for inference. This involves analyzing the unique characteristics of each cluster, such as purchasing patterns, product preferences, or consumer behavior. This information will provide deep insights into customer segmentation and help formulate more targeted business strategies. Data exploration on the clustering results may include creating visualizations, descriptive statistics, and additional modeling if needed. The main purpose of this exploration is to identify patterns or trends that can form the basis for strategic decision-making. Business strategy conclusions can be drawn after understanding the differences and similarities between customer groups.

The material or dataset used in this research is public data taken from the kaggle website. This research used the marketing_campaign.csv dataset with 29 attributes and 2240 rows. The experiment used Lenovo V14 laptop as client and a dedicated server with an NVIDIA Tesla P4 GPU under Google Colab. Python libraries such as NumPy, Pandas and Scikit Learn is used to facilitate data manipulation and model implementation. Jupyter Notebooks facilitates data exploration, visualization, and documentation of research results.

3 RESULTS AND DISCUSSION

This section shows the results of clustering using the Gaussian Mixture Model (GMM) algorithm. First, analyze the attributes that will be the focus of this clustering process such as: "MntWines", "MntFruits", "MntMeatProducts", "MntFishProducts", "MntSweetProducts", and "MntGoldProds". These attributes influence single important attribute called "Spending" which represents the total consumer expenditure. This process aims to summarize the contribution of each attribute to total spending, facilitating more effective analysis of consumer patterns and behaviors. This approach allows for more focused analysis on specific aspects of consumer behavior, providing deeper insights into preferences and spending trends within each category.

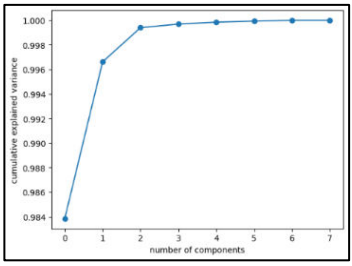


Fig. 1. Cumulative PCA variance

Next, using PCA to reduce dimensionality and improve efficiency. The PCA results can be seen in Figure 1. Components that are in PCA and have a high contribution will be applied to the GMM clustering algorithm. This result leads to determine the best number of clusters using the BIC algorithm. Figure 2 shows that the optimal parameters and number of clusters are using the diag parameter with the number of clusters 4. After clustering, the next step is to analyze and draw conclusions as follows:

a) Analysis of the Number of Clusters

Figure 2 shows the cluster data obtained from the calculation results using the Gaussian Mixture Model, there are 4 clusters. The proportion of customers is dominated by cluster 3 as much as 32%, followed by cluster 4 as much as 30% and cluster 2 as 27%, the rest are minority clusters 0 with value of 10%.

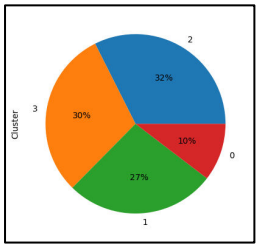


Fig. 2. Cluster Distribution

a) Expenditure and Income Analysis

Analytion result shows that based on the distribution of expenditure and income, cluster 3 tends to have the highest total expenditure compared to the amount of income, but the results of the analysis show that clusters 1 and 3 also have fairly high incomes but with less expenditure compared to other clusters.

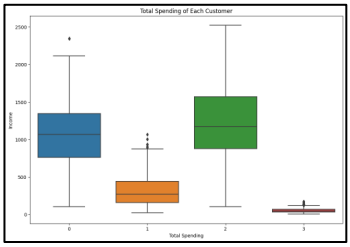
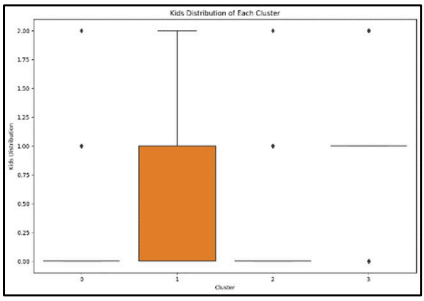


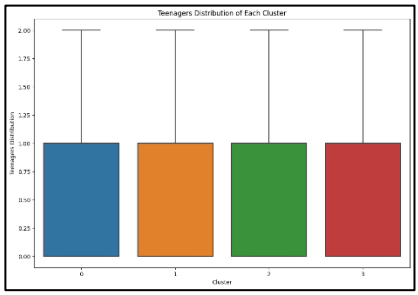
Fig. 3. Total Expenditure of Each Cluster

b) Marital Status Analysis

Another analysis results, shows that cluster 2 shows the most significant number of customers with the marital status of "married" and "living together" compared to other clusters. This position is followed by cluster 1 and cluster 3, which also show a few customers with similar marital status.



(a) Child distribution in each cluster



(b) Distribution of teenagers in each cluster

Fig. 3. *distribution in each cluster*

Figure 3.a illustrates that in cluster 1, only households with members from that group have children. Although cluster 2 has a diverse distribution of marital status and the largest number of customers, interestingly, none of them have children. On the other hand, Figure 3.b shows that all clusters including cluster 1 and cluster 2, have teenagers in their households. Figure 4 illustrates that each cluster shows almost similar product purchasing patterns, which means that so far, there is no significant difference in the number of items purchased by customers in each cluster. Several marketing strategies and business actions that can be taken are as follows:

a) Emphasis on Health and Active Lifestyle

Given that clusters 1 and 3 show a tendency to have significant spending on wine products, marketing strategies can be focused on promoting products that support a healthy and active lifestyle. Providing information related to the health benefits of wine consumption or even offering special offers for health products that are in line with the preferences of this cluster can increase customer appeal and satisfaction.

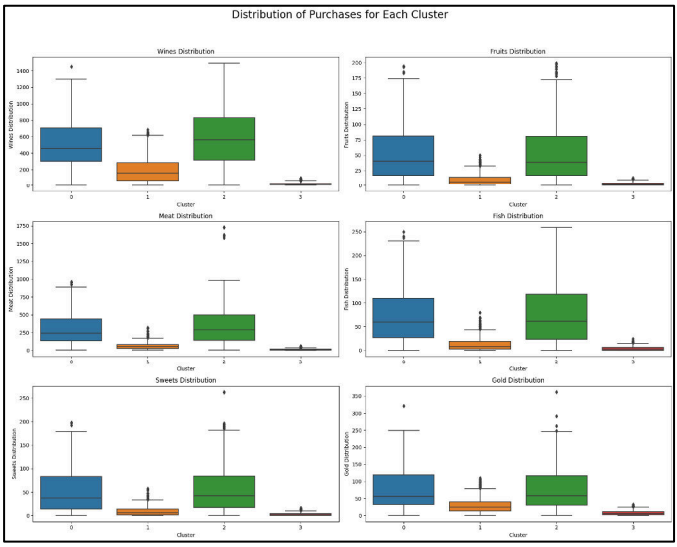


Fig. 4. Distribution of Customer Purchasing Products

b) Increasing Sales of Fruit, Meat and Fish Products:

Given that clusters 1 and 3 tend to buy a lot, marketing strategies can be focused on promoting fruit, meat, and fish products, including bundling offers or special discounts for these products. Providing packages that include various types of healthy food products can be an additional attraction for customers in this group. Thus, the company can take advantage of these consumer tendencies to increase sales of fruit, meat and fish products, while providing greater added value to customers.

c) Education for Low-Spending Clusters

For cluster 4, which tends to have lower spending, the focus of marketing strategies can be directed at educating customers about the value and benefits of the product, as well as special offers that are economically beneficial.

V. CONCLUSION

Based on the analysis that has been done, the following conclusions can be drawn. Dataset before clustering requires a data preprocessing stage and a feature selection stage is essential to ensure that the variables used in the analysis have a significant contribution. Selecting the right variables not only improves the accuracy of the clustering results but also helps avoid unexpected results in the cluster. In addition, the results of this analysis provide insight into the segmentation of buyer preferences and behavior. This data is expected to be used to design a more optimal and efficient business strategy. Further development of this case can be done by implementing the following suggestions. First, utilizing more diverse data, using data covering different geographic regions, can produce more representative and generally applicable rules. Second, considering external factors such as seasonality, market trends, or special events in the analysis process can provide additional perspectives on understanding buyer segmentation.

References

[1] G. Parkin, *Digital marketing: Strategies for online success*. Fox Chapel Publishing, 2016.
[2] D. Farahdiba, "Konsep dan strategi komunikasi pemasaran: perubahan perilaku konsumen menuju era disrupsi," *Jurnal Ilmiah Komunikasi Makna*, vol. 8, no. 1, pp. 22–38, 2020.
[3] A. Fadillah, "Perilaku Pembelian Pelanggan Ritel,". [Online]. Available: https://api.semanticscholar.org/CorpusID:240860110, 2019.
[4] A. Muh. Primabudi, "Analisa Faktor-Faktor Yang Mempengaruhi Keputusan Pembelian Pada Toko Online. Studi Kasus: Penjualan Game Secara Online," [Online]. Available: https://api.semanticscholar.org/CorpusID:198835004, 2017
[5] A. H. Massoudi, H. Q. Birdawod, and M. Raewf, "Personal Digital Marketing Influence on Successful Marketing Campaign in Today's Digital Age," *Cihan University-Erbil Journal of Humanities and Social Sciences*, [Online]. 2023.
[6] E. F. L. Awalina and W. I. Rahayu, "Optimalisasi Strategi Pemasaran dengan Segmentasi Pelanggan Menggunakan Penerapan K-Means Clustering pada Transaksi Online Retail," *Jurnal Teknologi dan Informasi*, 2023.
[7] C. Wang, "Efficient customer segmentation in digital marketing using deep learning with swarm intelligence approach," *Inf Process Manag*, vol. 59, no. 6, p. 103085, 2022.
[8] D. Suyanto, "Data Mining untuk klasifikasi dan klasterisasi data," *Bandung: Informatika Bandung*, 2017.
[9] Muh. N. Akbar, A. Salsabila, A. P. Asri, and M. Syawir, "ANALISIS CLUSTERING UNTUK SEGMENTASI PENGGUNA KARTU KREDIT DENGAN MENGGUNAKAN ALGORITMA K-MEANS DAN PRINCIPAL COMPONENT ANALYSIS," *AGENTS: Journal of Artificial Intelligence and Data Science*, 2023.
[10] A. Ivaschenko, A. Stolbova, and O. Golovnin, "Spatial clustering based on analysis of Big Data in digital marketing," in *Russian Conference on Artificial Intelligence*, Springer, pp. 335–347, 2019.
[11] F. Fatimah, F. Nataly, and Y. Purnamasari, "Penerapan Pemasaran Digital Dalam Meningkatkan Personal Branding," *Jurnal Abdimas Komunikasi dan Bahasa*, 2022
[12] A. A. Neath and J. E. Cavanaugh, "The Bayesian information criterion: background, derivation, and applications," *Wiley Interdiscip Rev Comput Stat*, vol. 4, no. 2, pp. 199–203, 2012.
[13] S. R. A. Ahmed, I. Al Barazanchi, Z. A. Jaaz, and H. R. Abdulshaheed, "Clustering algorithms subjected to K-mean and gaussian mixture model on multidimensional data set," *Periodicals of Engineering and Natural Sciences*, vol. 7, no. 2, pp. 448–457, 2019.